Using Discrete Tchebichef Transform on Speech Recognition

Ferda Ernawan, Edi Noersasongko[1] and Nur Azman Abu[2]
[1]Faculty of Information and Communication Technology Universitas Dian Nuswantoro
Semarang, Indonesia
[2]Faculty of Information and Communication Technology Universiti Teknikal Malaysia Melaka
Melaka, Malaysia
[1]e-mail: ferda1902@gmail.com, rektor@dinus.ac.id, [2]e-mail: nura@utem.edu.my

ABSTRACT

Speech recognition is becoming popular in current development on mobile devices. Presumably, mobile devices have limited computational power, memory size and battery life. In general, speech recognition is a heavy process that required large sample data within each window. Fast Fourier Transform (FFT) is the most popular transform in speech recognition. In addition, FFT operates in complex field with imaginary numbers. This paper proposes an approach based on discrete orthonormal Tchebichef polynomials as a possible alternative to FFT. Discrete Tchebichef Transform (DTT) shall be utilized here instead of FFT. The preliminary experimental result shows that speech recognition using DTT produces a simpler and efficient transformation for speech recognition. The frequency formants using FFT and DTT have been compared. The result showed that, they have produced relatively identical output in term of basic vowel and consonant recognition. DTT has the potential to provide simpler computing with DTT coefficient real numbers only than FFT on speech recognition.

Keyword-Speech Recognition; Fast Fourier Transforms; Discrete Tchebichef Transform.

1. INTRODUCTION

Transformation domain using FFT is widely used in speech recognition. The FFT is often used to compute numerical approximations to continuous Fourier. However, a straightforward application of the FFT to these problems often requires a large FFT computation to be performed even though most of the input data to this FFT may be zero. 1024 sample data FFT computation is considered the main basic algorithm for several digital signals processing. FFT algorithm is computationally complex transform which requires operating on an imaginary numbers. It is a complex exponential function that defines a complex sinusoid with frequency. The discrete Tchebichef transform is another transform method based on discrete Tchebichef polynomials [1]. DTT has a lower computational complexity and it does not require complex transform unlike continuous orthonormal transforms [2]. DTT does not involve any numerical approximation on friendly domain. The Tchebichef polynomials have unit weight and algebraic recurrence relations involving real coefficients. These factors in effect make DTT suitable for transforming the signal from time domain into frequency domain for speech recognition. In the previous work on DTT, it has been applied in several computer vision and image processing application. For examples, they are used in spectrum analysis of speech recognition [2], image reconstruction, image analysis [1] and image compression.

The organization of this paper is as follows. The definition of the discrete orthonormal Tchebichef polynomials is given in the next section. Section III presents the experimental result of speech recognition using FFT and DTT. The comparative of power spectral density, autoregressive model and frequency formants using FFT and DTT in speech recognition are discussed in Section IV and conclusion is in Section V.

2. DISCRETE ORTHONORMAL TCHEBICHEF POLYNOMIALS

The discrete orthonormal Tchebichef polynomials are proper especially when Tchebichef polynomials of large degree are required to be evaluated. For a given positive integer N (the vector size), and a value n in the range $[1,\ N-1]$, the orthonormal version of the one dimensional Tchebichef function is given by following recurrence relations in polynomials $t_k(n), n = 1,2,\dots,N-1$ [1]:

$$t_0(n) = \frac{1}{\sqrt{N}}, \tag{1}$$

Fourth International Conference on Machine Vision (ICMV 2011): Computer Vision and Image Analysis;
Pattern Recognition and Basic Technologies, edited by Zhu Zeng, Yuting Li, Proc. of SPIE Vol. 8350,
83501J · © 2012 SPIE · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.920206

$$t_k(0) = \sqrt{\frac{N-k}{N+k}}\sqrt{\frac{2k+1}{2k-1}}\,t_{k-1}(0), \tag{2}$$

$$t_k(1) = \left\{1 + \frac{k(1+k)}{1-N}\right\}t_k(0), \tag{3}$$

$$t_k(n) = \gamma_1 t_k(n-1) + \gamma_2 t_k(n-2), \tag{4}$$

$$k = 1, 2, \ldots, N-1,\ n = 2, 3, \ldots, \left(\frac{N}{2} - 1\right),$$

where

$$\gamma_1 = \frac{-k(k+1) - (2n-1)(n-N-1) - n}{n(N-n)}, \tag{5}$$

$$\gamma_2 = \frac{(n+1)(n-N-1)}{n(N-n)}, \tag{6}$$

The forward discrete Tchebichef transform of order N is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n)t_k(n), \tag{7}$$

$$k = 0, 1, \ldots, N-1,$$

where $X(k)$ denotes the coefficient of orthonormal Tchebichef polynomials. The inverse discrete Tchebichef transform is given by:

$$x(n) = \sum_{k=0}^{N-1} X(k)t_k(n), \tag{8}$$

$$n = 0, 1, \ldots, N-1,$$

The Tchebichef transform involves only algebraic expressions and it can be computed easily using a set of recurrence relations (1)-(6) above. The first five discrete orthonormal Tchebichef polynomials are shown in Fig. 1.

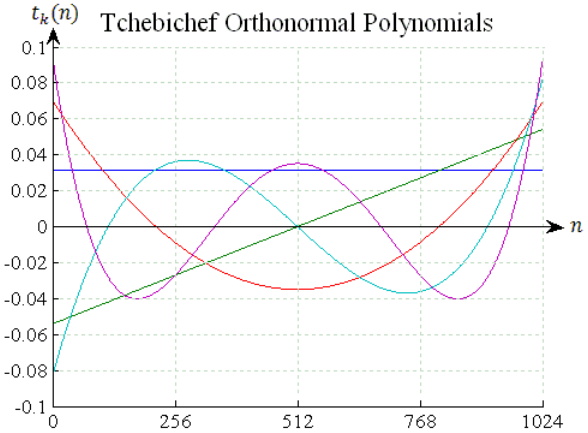

Figure 1. The First Five Discrete Orthonormal Tchebichef Polynomials $t_k(n)$ for $k = 0, 1, 2, 3$ and 4.

3. EXPERIMENTAL RESULTS

The sound of vowel ‘O’ shall be used is male voice from the Acoustic characteristics of American English vowels [3]. The sample sound of vowel 'O' has a sampling rate frequency component at about 10 KHz.

3.1 Windowing

Speech recognition using FFT technique used windowing function. A windowing function is used to smooth the discontinuities at the block boundary and then lessens the effect of spectral leakage. A window functions is used commonly in speech analysis to reduce the sudden changes and undesirable frequencies occurring in the framed speech. In the experiment, a windowing function is applied to computing the FFT. By applying the FFT method to finite duration sequences can produces inadequate result because of spectral leakage. Hamming window is given in the following equation:

$$w(k) = 0.54 - 0.46\cos\left[\frac{2\pi k}{L-1}\right] \tag{9}$$

where L represents the width of S_n and k is an integer, with values $0 \le k \le L-1$. The resulting windowed segment is defined as:

$$x(k) = S_n w(k) \tag{10}$$

where S_n is the signal function and $w(k)$is the window function. Whereas, DTT consist of only algebraic expressions and the Tchebichef polynomial matrix can be constructed easily using a set of recurrence relations. Therefore the window is very inefficient when the sample data are multiplied by a value close to zero. Any transition occurring during this part of the window will be lost so that the spectrum is no longer true real time. Speech recognition using DTT is not use windowing function. In this study, a sample speech signal has been windowed into 4 frames as illustrated in Fig. 2.

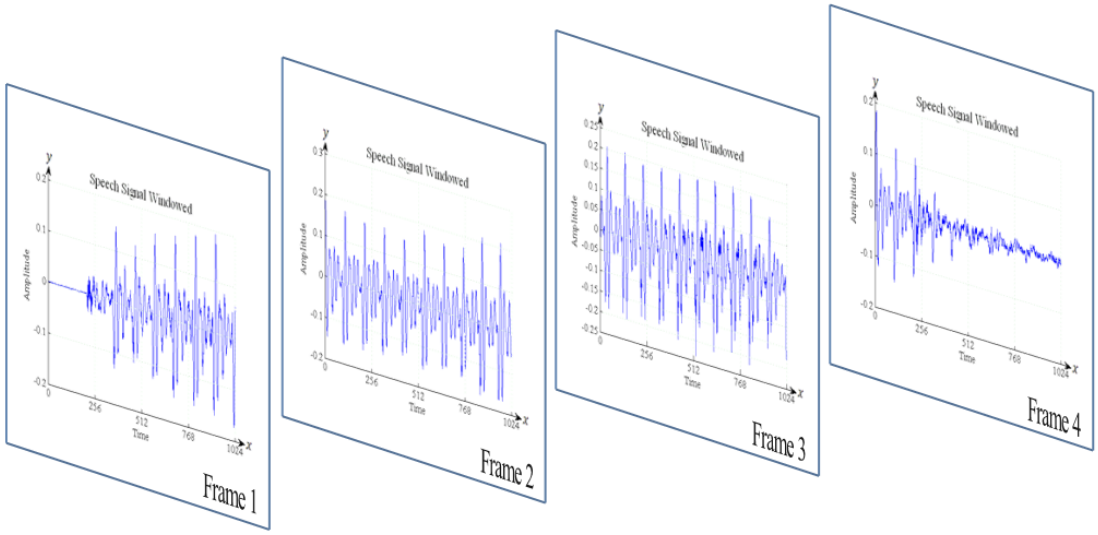

Figure 2. Speech signal windowed into four frames.

Each frame consume of 1024 sample data represent speech signal. In this study, the sample speech signal for 1-1024, 1025-2048, 2049-3072, 3073-4096 sample data which represent on frames 1, 2, 3, and 4. In this experiment, the sample speech signal on frame 4 is used to analyze and evaluate.

3.2 Coefficient of DTT

Next, speech signal on frame 4 is computed with 1024 discrete orthonormal Tchebichef polynomials. Coefficients of DTT of order n = 1024 sample data are given as follow formula:

$$\boldsymbol{TC} = \boldsymbol{S} \tag{11}$$

$$\begin{bmatrix} t_0(0) & t_0(1) & t_0(2) & \cdots & t_0(n-1) \\ t_1(0) & t_1(1) & t_1(2) & \cdots & t_1(n-1) \\ t_2(0) & t_2(1) & t_2(2) & \cdots & t_2(n-1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{n-1}(0) & t_{n-1}(1) & t_{n-1}(2) & \cdots & t_{n-1}(n-1) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-1} \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix}$$

where $\boldsymbol{C}$ is the coefficient of Discrete Tchebichef Transform, which represents $c_0, c_1, c_2, \dots, c_{n-1}$. $\boldsymbol{T}$ is matrix computation of discrete orthonormal Tchebichef polynomials $t_k(n)$ for $k = 0, 1, 2, \dots, N-1$. $\boldsymbol{S}$ is the sample of speech signal window which is given by $x(0), x(1), x(2), \dots, x(n-1)$. The coefficient of DTT is given in as follows:

$$C = T^{-1}S \tag{12}$$

3.3 Power Spectral Density

Power spectral density (PSD) showed the strength of the variations (energy) as a function of frequency. In other words, it shows at which frequencies variations are strong and at which frequencies variations are weak [4]. The one-sided PSD using DTT can be computed as:

$$pw(k) = 2\frac{|c(n)|^2}{(t_2 - t_1)} \tag{13}$$

where c(n) is the coefficient of discrete Tchebichef transform, the factor 2 is due to add for the contributions from positive and negative frequencies. (t_1, t_2) is precisely the average power of spectrum in the time range. The power spectral density is plotted using a decibel (dB) scale $20 \log 10$. The power spectral density using FFT and DTT for vowel 'O' on frame 4 is shown in Fig. 3.

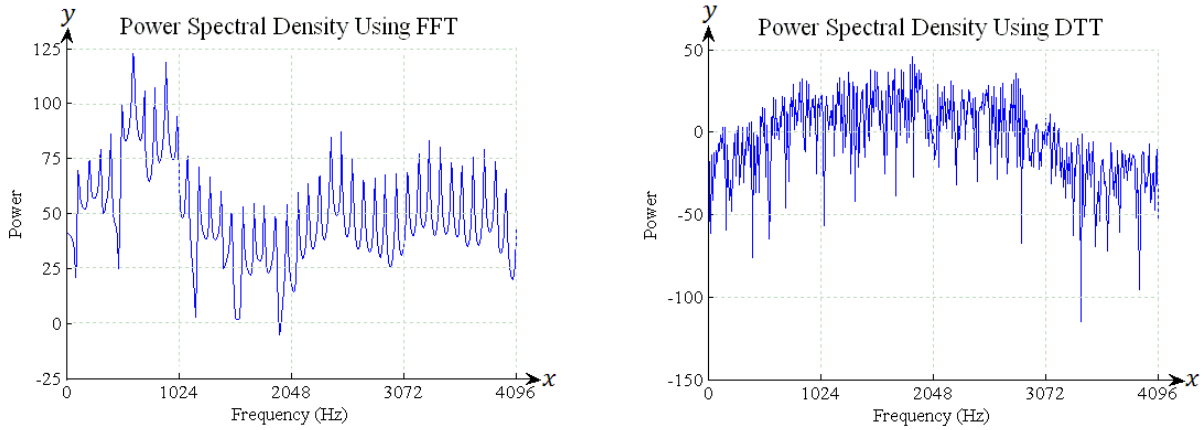

Figure 3. Power Spectral Density using FFT (left) and DTT (right) for vowel 'O' on frames 4.

3.4 Autoregressive

Autoregressive (AR) models are commonly obtained from the linier autocorrelation of discrete time signal to obtain an all pole estimate of the signal's power spectrum [5]. Autoregressive model is used to determine the characteristics of the vocal and to evaluate the formants. The autoregressive process of a series y_j using DTT of order v is given in the following equation:

$$y_j = -\sum_{k=1}^{v} a_k \, c_{j-k} + e_j \tag{14}$$

where a_k are real value autoregression coefficients, v is 12 and c_j is the coefficient of DTT at frequency index j. e_j represent the errors term independent of past samples. The autoregressive using FFT and DTT for vowel 'O' on frame 4 were shown in Fig. 4.

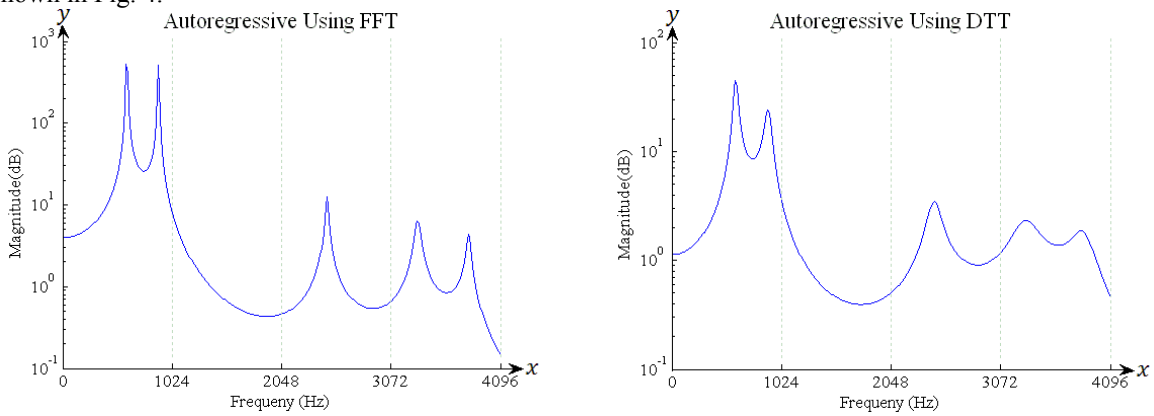

Figure 4. Autoregressive using FFT (left) and DTT (right) for vowel 'O' on frame 4.

3.5 Frequency Formants

Next, the frequency formant shall be detected. The uniqueness of each vowel is measured by formants. Formants are exactly the resonant frequencies of vocal tract [6]. A formant is a characteristic resonant region (peak) in the power spectral density of a sound. The formants of the autoregressive curve are found at the peaks using a numerical derivative. These vector positions of the formants are used to characterize a particular vowel. The Comparison of the frequency formants using FFT and DTT for the vowel 'O' on frame 4 is shown in Table I. The frequency peak formants of the experiment result F_1, F_2 and F_3 were compared to a referenced formants to decide on the output of the vowel. The referenced formants comparison code was written base on the classic study of vowels by Peterson and Barney [7].

Table 1. Frequency formants of vowels and consonants.

Vowel	Formants	FFT	DTT	Consonant	Formants	FFT	DTT
i	F_1	312	292	ka	F_1	796	721
	F_2	2265	2265		F_2	1152	1130
	F_3	3349	3300		F_3	2347	2336
ɛ	F_1	546	546	na	F_1	764	839
	F_2	1796	1777		F_2	1335	1345
	F_3	2470	2451		F_3	2519	2508
ɑ	F_1	712	703	pa	F_1	775	753
	F_2	1113	1103		F_2	1087	1065
	F_3	2480	2451		F_3	2573	2562
ɔ	F_1	605	595	ra	F_1	661	624
	F_2	908	898		F_2	1301	1248
	F_3	2480	2451		F_3	2160	2131
u	F_1	312	302	ta	F_1	829	796
	F_2	898	878		F_2	1162	1141
	F_3	2480	2451		F_3	2519	2530

4. DISCUSSION

In the sample above, an experimental result presented recognize vowels and consonants. The experiment result of speech recognition using FFT and DTT is compared and analyzed. The power spectral density of vowel 'O' using FFT on the left of Fig. 3 show that power spectrum is higher than power spectral density using DTT. It's because the result absolute square value of the speech signal using FFT consist of real part number and imaginary number. Next, the power spectral density using DTT produce more noise than FFT in frequency spectrum. According to observation as presented in the Fig. 4, the peaks of first frequency formant (F_1), second frequency formant (F_2) and third frequency formant (F_3) using FFT and DTT respectively was appear identically quite similar output. Based on the experiment result as presented in Table 1, the result of frequency formants of speech recognition using FFT and DTT for vowels and consonants respectively is nearly equally similar. Furthermore, speech recognition using DTT can be extended in the future to recognize a word or sentence from speech sound.

5. CONCLUSION

Speech recognition using FFT has been popular transform over the last decades. Alternatively, this paper introduces DTT on speech recognition. DTT produces a simpler and more computationally efficient than FFT. On the one hand, FFT is computationally complex with imaginary numbers. On the other hand, DTT consume simpler, faster computation with real coefficient number only. DTT is a potential contender as the next candidate to transform time domain into frequency domain. The autoregressive model using FFT and DTT produces the similar shape. The result showed that the peaks of vowel 'O' and consonant 'RA' using DTT were identically similar with FFT in term of vowel and consonant recognition. DTT has potential to perform well in term of basic vowel and consonant recognition.

REFERENCES

[1] R. Mukundan, "Some Computational Aspects of Discrete Orthonormal Moments," *IEEE Transactions on Image Processing*, Vol. 13, No. 8, Aug. 2004, pp. 1055-1059.

[2] F. Ernawan and N.A. Abu "Efficient Discrete Tchebichef on Spectrum Analysis of Speech Recognition," *International Journal of Machine Learning and Computing*, Vol. 1, No. 1, Apr. 2011, pp. 1-6.

[3] J. Hillenbrand, L.A. Getty, M.J. Clark, and K. Wheeler, "Acoustic characteristic of American English vowels," *Journal of the Acoustical Society of America*, Vol. 97, No. 5, May 1995, pp. 3099-3111.

[4] A.H. Khandoker, C.K. Karmakar and M. Palaniswami, "Power spectral analysis for identifying the onset and termination of obstructive sleep apnoea events in ECG recordings," *Proceeding of the 5th International Conference on Electrical and Computer Engineering (ICECE 2008)*, Dec. 2008, pp. 096-100.

[5] M. Athineos, and D.P.W. Ellis, "Autoregressive Modeling of Temporal Envelopes," *IEEE Transaction on Signal Processing*, Vol. 55, No. 11, Nov. 2007, pp. 5237-5245.

[6] A. Patil, C. Gupta, and P. Rao, "Evaluating Vowel Pronunciation Quality: Formant Space Matching Versus ASR Confidence Scoring," *National Conference on Communication (NCC)*, Jan. 2010, pp. 1-5.

[7] G.E. Peterson, and H.L. Barney, "Control Methods Used in a Study of the Vowels," *Journal of the Acoustical Society of America*, Vol. 24, No. 2, Mar. 1952, pp. 175-184.